Journal of Behavioral and Experimental Economics 112 (2024) 102273

Contents lists available at ScienceDirect

Journal of Behavioral and Experimental Economics

journal homepage: www.elsevier.com/locate/jbee

Manipulating response times in the cognitive reflection test: Time delay boosts deliberation, time pressure hinders it

Ennio Bilancini [a], Leonardo Boncinelli [b], Tatiana Celadin [c,*]

[a] MT School for Advanced Studies, Piazza S. Francesco 19, 55100 Lucca, Italy
[b] Department of Economics and Management, University of Florence, Via delle Pandette 9, 50127 Firenze, Italy
[c] Department of Economics, Ca' Foscari University of Venice, Cannaregio 873, Fondamenta S.Giobbe, 30121 Venezia, Italy

ARTICLE INFO

JEL classification:
D91

Keywords:
Cognitive reflection test
Time pressure
Time delay

ABSTRACT

We employ the Cognitive Reflection Test (CRT) to assess the effectiveness of two widely applied cognitive manipulations which rely on constraining response times. The CRT measures the ability of a person to resist giving an immediate response that is intuitive but incorrect in favor of greater reflection to find out the correct answer. We have two treatments: a Time Pressure (TP) treatment (provide an answer within 30 s) and a Time Delay (TD) treatment (provide an answer after 60 s). We find that TD increases the frequency of correct answers, while TP increases the frequency of incorrect answers, especially incorrect answers that are not intuitive. Moreover, we confirm the existence of gender biases as already found in other studies. In particular, gender moderates the effects of the experimental treatments: TD increases the frequency of correct answers for males but not for females, while TP increases the frequency of incorrect but less intuitive answers for females but not for males. Our findings provide important insights on the effectiveness of the time manipulations that are widely used in the literature of cognition.

1. Introduction

Dual-Process Theories, which posit that decision-making is affected by the interplay between two cognitive processes (Evans & Stanovich, 2013), have become popular in the study of humans' reasoning, judgment, and social behavior (Evans, 1989; Wason & Evans, 1974). According to these theories, there are two modes of cognition: one is rapid and autonomous, it produces default responses, and is associated with intuition; the other one requires greater reasoning, larger use of working memory, cognitive decoupling, mental simulation, and it is slow, controlled, and associated with deliberation (Evans & Stanovich, 2013). Dual-process theories have been broadly applied, among other things, to the study of cooperation (Alós-Ferrer & Garagnani, 2020; Bilancini et al., 2022; Capraro & Cococcioni, 2015, 2016; Lohse, 2016; Rand et al., 2012, 2014), honesty (Capraro, 2017; Capraro, Schulz, & Rand, 2019; Gunia et al., 2012; Lohse et al., 2018), deontology and utilitarianism (Cummins & Cummins, 2012; Suter & Hertwig, 2011; Trémolière & Bonnefon, 2014).

To investigate the causal effect of the modes of cognition on the aforementioned behaviors, several experimental treatments have been developed with the purpose of promoting reliance on intuition or deliberation. An important class is given by the experimental treatments relying on manipulating response times. Among these, Time Pressure (TP) and Time Delay (TD) have been widely applied in experimental works (Alós-Ferrer & Garagnani, 2020; Bilancini et al., 2022; Rand et al., 2012, 2014). In the TP treatment, individuals are forced to answer within a short amount of time, to increase reliance on intuition. In the TD treatment, individuals are forced to answer after a certain amount of time, to increase reliance on deliberation. Manipulating response time has been widely used in psychology (in particular in cognitive psychology, Capraro 2024), and there is broad agreement that the constraint of response time tends to reduce reliance on deliberation; however, it is less clear to what extent it promotes reliance on intuition. An important aspect is the actual time span considered. In the cognitive psychology literature, the time span is of the magnitude of hundreds of milliseconds, to focus on automatic and unconscious responses. In this paper, we consider a time span of the magnitude of seconds, to focus on intuitive responses that are conscious and based on heuristics (Belloc et al., 2019; Capraro, 2024). Although many studies have employed these two treatments to explore how intuition and deliberation may affect behavior, so far no direct behavioral test has been done to see if these treatments actually do what they have been designed for.

In this paper, we test the effectiveness of TP and TD using the Cognitive Reflection Test (CRT, Frederick 2005), which is a measure of the ability or the disposition of a person to engage in a more deliberative

[*] Corresponding author.
E-mail addresses: ennio.bilancini@imtlucca.it (E. Bilancini), leonardo.boncinelli@unifi.it (L. Boncinelli), tatiana.celadin@unive.it (T. Celadin).

https://doi.org/10.1016/j.socec.2024.102273
Received 25 July 2023; Received in revised form 18 July 2024; Accepted 19 July 2024
Available online 24 July 2024

process and to resist intuitive responses (Branas-Garza et al., 2019). We use the responses provided to the CRT as a proxy of the effectiveness of TP and TD. The CRT was originally designed to capture the disposition to be reflective, not as a measure of actual reflection (Frederick, 2005). Specifically, the questions developed for the CRT aim at prompting an intuitive but incorrect answer, with deliberation taking the form of resisting the intuitive response to engage in further reflection. So, it seems reasonable to expect that experimental treatments designed to manipulate reliance on intuition and deliberation would have an effect on the answers provided to the CRT. Another feature of the CRT is that the resulting score is a proxy of physiological characteristics (Alonso et al., 2018; Bosch-Domènech et al., 2014), and it has the property to predict individuals' performance, decision-makers' choices, and behaviors (Albano et al., 2018; Andersson et al., 2016; Besedeš et al., 2012; Campitelli & Labollita, 2010; Frederick, 2005; Ponti & Rodriguez-Lara, 2015).

We find that the TD treatment increases the likelihood of providing correct answers to the CRT and this suggests that the TD treatment is effective in promoting reliance on deliberation. Moreover, the TP treatment does not increase the likelihood of providing the intuitive but incorrect answers that the CRT is designed to induce, while it increases the likelihood of providing non-intuitive incorrect answers. This suggests that the TP treatment is effective in reducing reliance on deliberation (results are even stronger when we look only at the new version of the CRT). In the literature, the intuitive incorrect answers provided to the CRT have been used as a measure of the disposition to behave intuitively (Cueva et al., 2016), although its reliability has been criticized. There are two possible explanations for why TP has such an effect. The first possibility is that TP does not foster intuition much but it mostly impairs correct reasoning, leading to confusion and thus a greater likelihood of wrong (random) answers rather than intuitive ones (Goeschl & Lohse, 2018). The other possibility is that the intuitive answers of the CRT do not really capture intuition. Indeed, there is evidence that while correct answers in the CRT are a reliable measure of deliberation, intuitive answers may not be a reliable measure of intuition (Pennycook et al., 2016).

Our results provide empirical support for the effectiveness of cognitive manipulations based on experimental treatments imposing constraints on response times. Specifically, we provide evidence that Time Delay increases reliance on deliberation and Time Pressure decreases reliance on intuition, and this is an important result for the literature in Dual Process Theories. Furthermore, we provide a novel approach that can be used to test the effectiveness of other cognitive manipulations, possibly expanding the scope of the CRT.

We stress that such results rely on the presumption that CRT is an effective measure of the ability or the disposition of a person to engage in a more deliberative process and to resist intuitive responses. Alternatively, one could see our experiment as a test of the effectiveness of the CRT by assuming that TP and TD are effective cognitive manipulations.

2. Methods

We recruited 598 participants using the online platform Prolific [(Palan & Schitter, 2018), www.prolific.co].

Participants were randomly assigned to one of three treatments: a Baseline, Time Pressure treatment, and Time Delay treatment. In each treatment, participants had to answer the six questions of the CRT-L (Primi et al., 2016) that were presented in random order. Since people on Prolific may have seen the original CRT several times, it is possible that they already know the answers and respond by direct recall. For this reason, we opted to use a longer version of the CRT, which includes three new questions that are less frequently used.

In the Baseline, participants had to answer each question without any time constraints. In the TP treatment, participants had to answer each question within 30 s (Borghans et al., 2008). Participants who failed to answer within the time constraint were still able to provide the answer to the questions. Indeed, after 30 s the question was still shown on the screen and participants were still able to provide an answer. Overall 94.69% of participants were able to answer within the time constraint (in the Appendix we report the percentage of compliance for each question, and we rerun the main analysis restricted to those participants who were able to provide an answer within 30 s). In the TD treatment, participants had to wait for one minute before they could insert an answer (Borghans et al., 2008); only after this amount of time participants were allowed to provide an answer. At the end of the study, we asked participants to assess their level of reflection in answering the CRT-L. Moreover, we elicited participants' demographic information as well as previous exposure to the CRT-L (instructions in the Appendix). The distribution of individuals' characteristics, such as gender, age, education, previous experience, and student and employment status are well balanced across treatments. The participation fee was 0.63 GBP for a survey 6:46 min long in mean (average reward per hour: 7.49 GBP). Since participants were not incentivized for the answers provided to the CRT-L, we acknowledge the possibility that we have been selecting a pool of participants that are diligent in solving the tasks (see Goeschl & Lohse, 2018, on this). The design, the analysis, the sample size, and the exclusion criteria were pre-registered at AsPredicted.org and can be found at this permanent address: https://aspredicted.org/49X_GRT.

3. Results at question level

We pre-registered a sample size of N=600. This was based on an a priori power analysis that showed that 200 subjects per treatment are enough to detect an effect size of f=0.25 with alpha=0.05 and power 0.80. After downloading the data file from Qualtrics, we obtained 598 observations (the server failed at registering two observations). We collected 207 participants in the Baseline, 204 in the TP treatment and 187 in the TD treatment.

Following our pre-registration, we first make an overall comparison using Kruskal–Wallis at the question level to test the difference in the distributions of the correct answers to each question of the CRT-L across all treatments. Specifically, our main variable is a dummy variable that takes value = 1 if the answer is correct, 0 otherwise. We find that the distribution differs significantly across treatments (Kruskal–Wallis test: χ^2 = 36.343, p<0.001).

The pairwise comparisons of the correct answers between treatments are statistically significant (see Fig. 1a). The likelihood of providing correct answers is statistically lower in the TP treatment compared to the Baseline (Wilcoxon rank-sum test: z = 3.167, p=0.002), while it is statistically higher in the TD treatment compared to the Baseline (Wilcoxon rank-sum test: z = −2.959, p=0.003), and higher in the TD treatment compared to the TP treatment (Wilcoxon rank-sum test: z = −6.027, p<0.001). In Table 1 we run Logit regressions with standard errors clustered at the individual level. Model 1 in Table 1 confirms that the TP treatment decreases the probability of providing a correct answer to the CRT, while TD increases the probability. The direction of these effects is confirmed when we control for gender, previous exposure to the CRT, question and order fixed effects (Model 2), though with somewhat larger standard errors.

We now consider the distribution of the intuitive answers (main variable=1 if individuals provided an intuitive response, 0 otherwise), and we find that the distributions are marginally different (Kruskal–Wallis tests: χ^2 = 6.183, p=0.045). The pairwise comparisons between treatments are statistically significant, with the exception of the Baseline and the TP treatment (Wilcoxon rank-sum test Baseline vs. TP: z = −0.277, p=0.782; Wilcoxon rank-sum test Baseline vs. TD: z = 2.031, p=0.042; Wilcoxon rank-sum test TD vs. TP: z = 2.294, p=0.022; see Fig. 1b). Models 3 and 4 in Table 1 show that the TP and TD treatments do not affect the probability of providing an intuitive answer to the CRT.

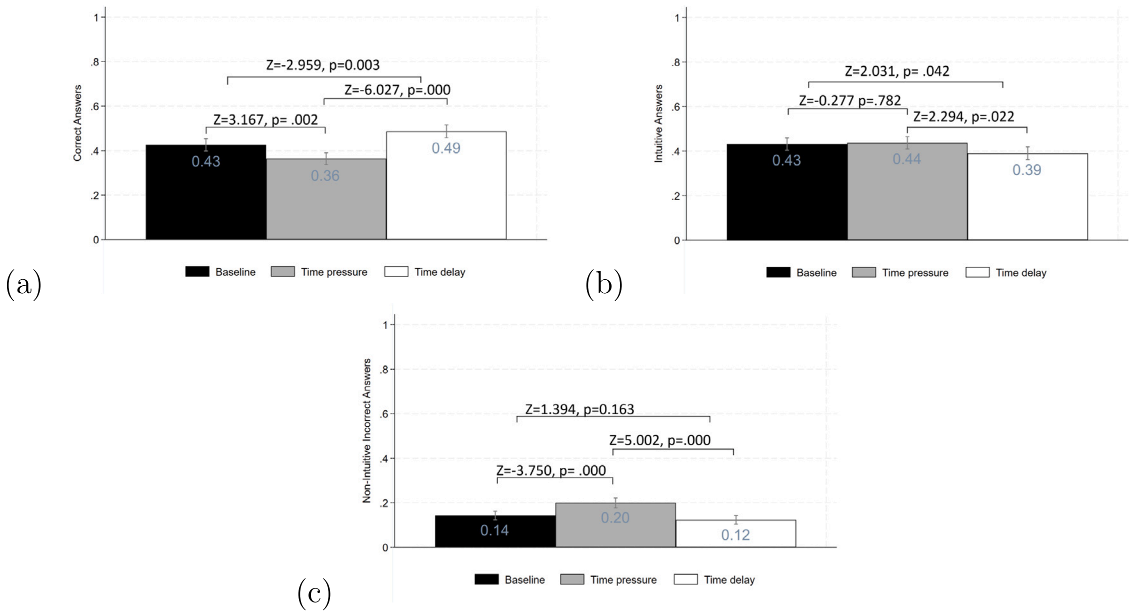

Fig. 1. (a) The mean of the correct answers provided to the CRT-L across treatments. (b) The mean of the intuitive answers provided to the CRT-L across treatments. (c) The mean of the non-intuitive incorrect answers provided to the CRT-L across treatments.

Table 1

Logit Regression on the likelihood of providing correct, intuitive, and non-intuitive incorrect answers to the CRT-L. *Correct* = 1 if the answer is correct, 0 otherwise; *Intuitive* = 1 if the answer is intuitive, 0 otherwise; *Non-Intuitive Incorrect* = 1 if the answer is non-intuitive incorrect, 0 otherwise; *TD* = 1 if a participant is under Time Delay, 0 otherwise; *TP* = 1 if a participant is under Time Pressure, 0 otherwise; *Female* = 1 if female, 0 otherwise; *Exposure* = 1 if individuals have seen someone of the CRT-L questions or all of the CRT-L questions, 0 if individuals have seen none of the CRT-L questions. *No Compliance* = 1 if a participant did not comply with the time manipulation, 0 otherwise. Robust standard errors in parentheses clustered at the individual level.

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
					Non-Intuitive	Non-Intuitive
	Correct	Correct	Intuitive	Intuitive	Incorrect	Incorrect
TP	−0.261[a]	−0.264	0.023	0.054	0.404[b]	0.334[b]
	(0.134)	(0.163)	(0.118)	(0.133)	(0.132)	(0.153)
TD	0.245[a]	0.282[a]	−0.170	−0.178	−0.170	−0.191
	(0.135)	(0.149)	(0.119)	(0.124)	(0.141)	(0.150)
Female		−0.746[c]		0.543[c]		0.296[b]
		(0.122)		(0.101)		(0.119)
Exposure		0.554[c]		−0.564[c]		0.045
		(0.137)		(0.121)		(0.139)
No Compliance		0.025		−0.270		0.362[a]
		(0.231)		(0.205)		(0.220)
Question	No	Yes	No	Yes	No	Yes
Order	No	Yes	No	Yes	No	Yes
Constant	−0.298[b]	−0.300[a]	−0.275[c]	−0.326[b]	−1.795[c]	−1.749[c]
	(0.095)	(0.157)	(0.084)	(0.150)	(0.095)	(0.186)
N	3588	3588	3588	3588	3588	3588
pseudo R^2	0.007	0.112	0.001	0.067	0.009	0.087

[a] Denotes $p < 0.10$.

[b] Denotes $p < 0.05$.

[c] Denotes $p < 0.01$.

Finally, when we consider the distribution of the non-intuitive incorrect answers (main variable=1 if individuals provided a non-intuitive incorrect response, 0 otherwise) we find that the distributions differ significantly across treatments (Kruskal–Wallis tests: $\chi^2 = 28.487$, p<0.001). The pairwise comparisons between each treatment are statistically significant, with the exception between the Baseline and the TD treatment (Wilcoxon rank-sum test Baseline vs. TP: $z = -3.750$, p<0.001; Wilcoxon rank-sum test Baseline vs. TD: $z = 1.394$, p=0.163; Wilcoxon rank-sum test TD vs. TP: $z = 5.002$, p<0.001; see Fig. 1c). Model 5 in Table 1 shows that TP increases the probability of providing non-intuitive incorrect answers to the CRT, while TD has no effect. Results hold even when we control for gender, previous exposure, question and order fixed effects (Model 6).

We report the Chi-squared test to test whether correct, intuitive, and non-correct intuitive responses differ significantly across treatments, and we confirm previous results (correct: $\chi^2_{(2)} = 36.353$, p<0.001; intuitive: $\chi^2_{(2)} = 6.185$, p=0.045; non-intuitive incorrect: $\chi^2_{(2)} = 28.495$, p<0.001). We run a Fisher's exact test to make the pairwise comparison, and we confirm previous results (see Table 2).

4. Results at individual level

Our second pre-registered variables are the number of correct, intuitive and non-intuitive incorrect answers provided to the CRT-L at the individual level. Following our pre-registration, we first make an overall comparison using Kruskal–Wallis to test differences in the

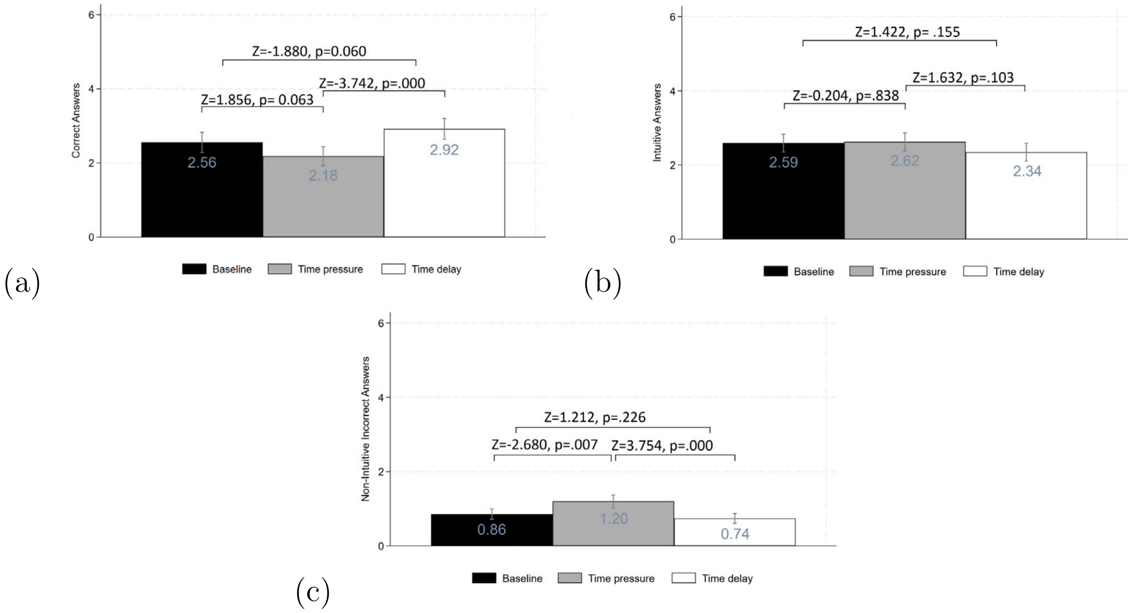

Fig. 2. (a) The mean of the correct answers provided to the CRT-L across treatments. (b) The mean of the intuitive answers provided to the CRT-L across treatments. (c) The mean of the non-intuitive incorrect answers provided to the CRT-L across treatments.

Table 2
Fisher's exact test. BL=Baseline; TP= Time Pressure and TD= Time Delay.

	Fisher's exact test		
	BL-TP	BL-TD	TP-TD
Correct	$p = 0.002$	$p = 0.003$	$p < 0.001$
Intuitive	$p = 0.807$	$p = 0.045$	$p = 0.023$
Non-intuitive incorrect	$p < 0.001$	$p = 0.182$	$p < 0.001$

distributions across all treatments. The Kruskal–Wallis test on the distribution of the number of correct answers finds statistically significant differences across treatments (χ^2 = 13.858, p-value = 0.001). The Kruskal–Wallis test on the distribution of the number of intuitive answers does not find any statistically significant difference across treatments (χ^2 = 3.103, p-value = 0.212). The Kruskal–Wallis test on the distribution of the number of non-intuitive incorrect answers finds statistically significant differences across treatments (χ^2 = 15.243, p-value = 0.001). In the Appendix, we provide detailed information on the distributions and the Tobit regressions (see Table 8).

Looking at the pairwise comparisons, we find that the number of correct answers provided to the Baseline and TP treatment shows a marginal difference (Wilcoxon rank-sum test: z = 1.856, p-value = 0.063). The difference between the Baseline and TD treatment is also marginally significant (Wilcoxon rank-sum test: z = –1.880, p-value = 0.060), while the difference between TP and TD is statistically significant (Wilcoxon rank-sum test: z = –3.742, p-value <0.001), see Fig. 2a.

The pairwise comparisons between the number of intuitive answers do not reveal a statistically significant difference across treatments (all p > 0.1, see Fig. 2b). Indeed, there is no significant difference in the number of intuitive answers provided to the Baseline and TP treatment (Wilcoxon rank-sum test: z = –0.204, p-value = 0.838), in the Baseline and TD treatment (Wilcoxon rank-sum test: z = 1.422, p-value = 0.155), and in the TP and TD treatment (Wilcoxon rank-sum test: z = 1.632, p-value = 0.103).

We find statistically significant differences in the pairwise comparisons between treatments for the non-intuitive incorrect answers (see Fig. 2c) except for the comparison between the Baseline and the TD treatment (Wilcoxon rank-sum test Baseline vs TP: z = –2.680, p-value = 0.007; Wilcoxon rank-sum test Baseline vs TD: z = 1.212, p-value = 0.226; Wilcoxon rank-sum test TD vs TP: z = 3.754, p-value <0.001).

5. Exploring the role of gender

There is large consensus in the literature that CRT exhibits a significant gender difference, with males performing better than females (Branas-Garza et al., 2019; Cueva et al., 2016; Frederick, 2005; Holt et al., 2017; Hoppe & Kusterer, 2011). Therefore, beyond the pre-registered analysis, we explore the effects of cognitive manipulations on CRT-L performance separately for males and females, both at the question and at the individual level.

Starting with the results at the question level, we have that the likelihood of providing correct answers is higher for males than females and this is confirmed by Model 1 in Table 3, and our results are consistent with the literature. Additionally, when examining the effects of the two treatments on males and females separately, we observe that males have a higher likelihood of providing correct answers under the TD treatment compared to both the Baseline and TP treatment. Under TP the likelihood is lower than the Baseline (see Fig. 3a). Results are confirmed by Model 3 in Table 3. For females, there is no significant difference in the likelihood of providing correct answers between the TD treatment and the Baseline. However, TP decreases the likelihood compared to both the Baseline and the TD treatments (see Fig. 3a), although this is not confirmed by Model 2 in Table 3. Overall, results suggest that males are more responsive to cognitive manipulations than females and exposure to TD increases the likelihood of providing correct answers for males but not for females. Wilcoxon rank-sum tests are in Fig. 3a.

The likelihood of providing intuitive answers is higher for females than males as confirmed by Model 4 in Table 3, and our results are consistent with the literature. When examining the effect of the two treatments on males and females separately, we observe that for males the likelihood of providing intuitive answers is lower under TD compared to the Baseline and TP. There is no significant difference in the likelihood of providing intuitive answers between the Baseline and TP treatment (see Fig. 3b). Results are confirmed by Model 6 in Table 3. For females, there is no significant difference in the likelihood of providing intuitive answers across treatments (Fig. 3b). Results are confirmed by Model 5 in Table 3. These findings suggest that males

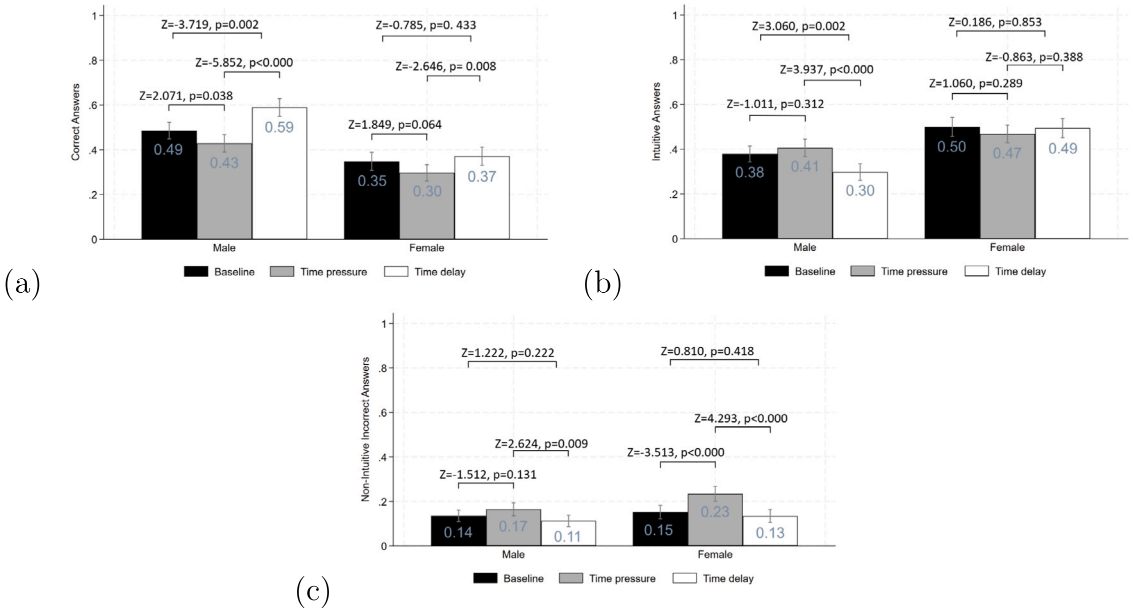

Fig. 3. (a) The mean of the correct answers provided to the CRT-L across treatments and gender. (b) The mean of the intuitive answers provided to the CRT-L across treatments and gender. (c) The mean of the non-intuitive incorrect answers provided to the CRT-L across treatments and gender.

are more responsive to cognitive manipulations than females, while exposure to CRT decreases the likelihood of providing correct answers for both males and females. Wilcoxon rank-sum tests are in Fig. 3b.

Finally, the likelihood of providing non-intuitive incorrect answers is the same for females and males, as confirmed by Model 7 in Table 3. When examining the effect of the two treatments on males and females separately, we observe that for males the likelihood of providing non-intuitive incorrect answers is lower under TD compared to both the Baseline and TP treatment (see Fig. 3c). Results are confirmed by Model 9 in Table 3. For females, TP increases the likelihood of providing non-intuitive incorrect answers (Fig. 3c). Results are confirmed by Model 8 in Table 3. Wilcoxon rank-sum tests are in Fig. 3c.

We now examine the effect of cognitive manipulations on males and females looking at the answers at the individual level.

Overall, males tend to provide a higher number of correct answers compared to females (see Table 9 in the Appendix), and this is consistent with the literature. When examining the effects of the two treatments on males and females separately, we observe that males provide a higher number of correct answers under TD with respect to TP and the Baseline (see Fig. 4a). For females, the number of correct answers is marginally higher under TD with respect to TP, but overall there is no difference across treatments (see Fig. 4a). Overall, it appears that males are more responsive to cognitive manipulations compared to females. Additionally, exposure to CRT increases the likelihood of providing correct answers for females, while this effect is not observed in males. Wilcoxon rank-sum tests are in Fig. 4a.

We now examine the intuitive answers and, our findings indicate that females tend to provide a higher number of intuitive answers compared to males, and our results are consistent with the literature. When examining the effect of the two treatments on males and females separately, we observe that for males under TD, the number of intuitive answers is lower with respect to TP and the Baseline (see Fig. 4b). For females, there is no difference in the number of intuitive answers across treatments (Fig. 4b). Once again, it appears that males exhibit greater responsiveness to cognitive manipulations compared to females. Furthermore, exposure to CRT appears to decrease the likelihood of providing correct answers for both males and females. Wilcoxon rank-sum tests are in Fig. 4b.

Finally, females and males provide the same number of non-intuitive incorrect answers. When examining the effect of the two treatments on males and females separately, we observe that males under TD provide a slightly lower number of non-intuitive incorrect answers with respect to TP and the Baseline (see Fig. 4c). For females under TP, the number of non-intuitive incorrect answers is higher compared to TD and the Baseline (Fig. 4c). Wilcoxon rank-sum tests are in Fig. 4c.

6. Discussion

In this paper, we have studied the effectiveness of two widely used experimental treatments based on constraining response times, Time Pressure (TP) and Time Delay (TD). Our novelty lies in using the answers provided to the Cognitive Reflection Test as a measure of their effectiveness (Frederick, 2005).

Our data show that the TD treatment increases the frequency of correct answers to the CRT, suggesting that the TD treatment is effective in promoting reliance on deliberation. Further, the TP treatment increases the frequency of incorrect answers, suggesting that the TP treatment is effective in reducing reliance on deliberation. These results are even stronger when we restrict our analysis to the new version of the CRT by Primi et al. (2016), (see Table 13). Indeed, TD increases the likelihood of providing correct answers, while it decreases the likelihood of providing intuitive answers. While TP decreases the likelihood of providing correct answers and increases the likelihood of providing non-intuitive incorrect answers. This is very important because the version of the CRT by Frederick (2005) has been seen many times, and people might answer correctly to the answers because they know the responses ex ante. In contrast, the new version by Primi et al. (2016) allows us to confirm our results.

Our data also confirm previous results on gender bias regarding CRT scores: females were more likely to provide intuitive responses and less likely to provide correct answers, compared to males. Interestingly, we also find that the effect of the experimental treatments is gender-specific: TD increases the frequency of correct answers for males but not for females, while TP increases the frequency of non-intuitive incorrect answers for females but not for males. Moreover, females tend to perform better the more they have been exposed to CRT questions in the past.

It seems natural to ask why the TP treatment may have these effects. Our results are compatible with at least two distinct interpretations. One possibility is that TP does not foster intuition much but instead

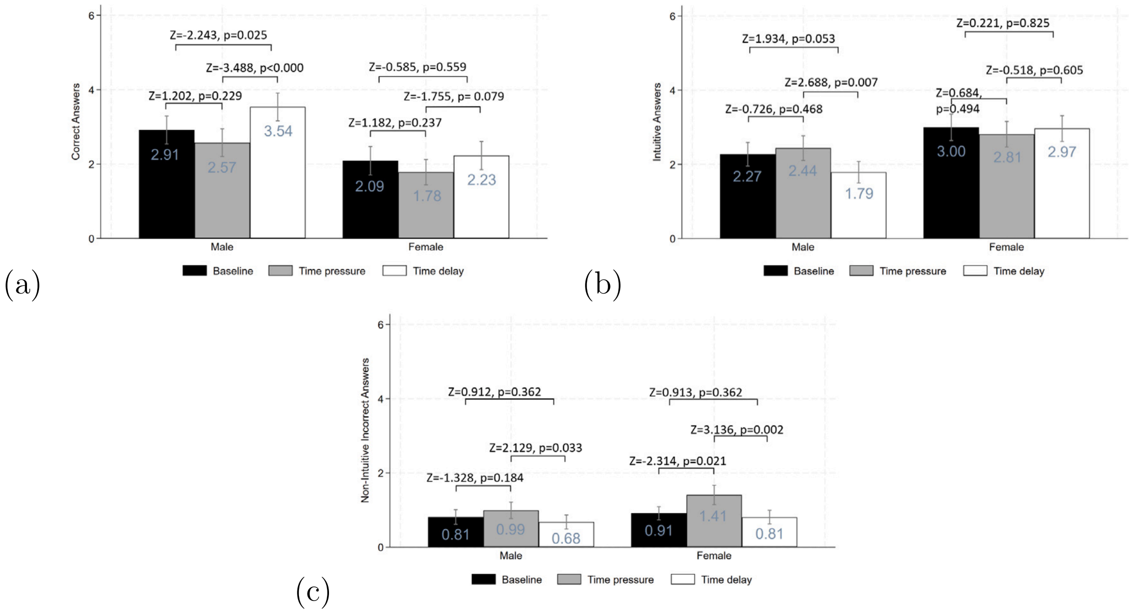

Fig. 4. (a) The mean of the correct answers provided to the CRT-L across treatments and gender. (b) The mean of the intuitive answers provided to the CRT-L across treatments and gender. (c) The mean of the non-intuitive incorrect answers provided to the CRT-L across treatments and gender.

Table 3
Logit Regression on the likelihood of providing correct, intuitive, and non-intuitive incorrect answers to the CRT-L. *Correct* = 1 if the answer is correct, 0 otherwise; *Intuitive* = 1 if the answer is intuitive, 0 otherwise; *Non-Intuitive Incorrect* = 1 if the answer is non-intuitive incorrect, 0 otherwise; *TD* = 1 if a participant is under Time Delay, 0 otherwise; *TP* = 1 if a participant is under Time Pressure, 0 otherwise; *Female* = 1 if female, 0 otherwise; *Exposure* = 1 if individuals have seen someone of the CRT-L questions or all of the CRT-L questions, 0 if individuals have seen none of the CRT-L questions. *No Compliance* = 1 if a participant did not comply with the time manipulation, 0 otherwise. Robust standard errors in parentheses clustered at the individual level.

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9
							Non-Intuitive	Non-Intuitive	Non-Intuitive
	Correct	Correct	Correct	Intuitive	Intuitive	Intuitive	Incorrect	Incorrect	Incorrect
TP	−0.252	−0.226	−0.312	0.161	−0.098	0.204	0.184	0.473[b]	0.199
	(0.212)	(0.235)	(0.217)	(0.178)	(0.193)	(0.181)	(0.222)	(0.201)	(0.230)
TD	0.439[b]	0.068	0.449[b]	−0.359[b]	0.017	−0.367[b]	−0.221	−0.153	−0.229
	(0.203)	(0.224)	(0.201)	(0.174)	(0.184)	(0.173)	(0.229)	(0.193)	(0.229)
Female	−0.619[c]			0.508[c]			0.154		
	(0.211)			(0.175)			(0.199)		
TP×Female	−0.036			−0.218			0.302		
	(0.297)			(0.247)			(0.279)		
TD×Female	−0.355			0.364			0.071		
	(0.297)			(0.248)			(0.298)		
Exposure	0.552[c]	0.889[c]	0.298	−0.560[c]	−0.818[c]	−0.341[b]	0.040	0.025	0.063
	(0.137)	(0.200)	(0.183)	(0.121)	(0.180)	(0.162)	(0.139)	(0.181)	(0.210)
No Compliance	0.017	−0.323	0.310	−0.249	−0.061	−0.466	0.350	0.413	0.248
	(0.232)	(0.348)	(0.328)	(0.205)	(0.258)	(0.345)	(0.219)	(0.312)	(0.292)
Question	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Order	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Constant	−0.351[b]	−1.165[c]	−0.209	−0.312[a]	0.209	−0.340[a]	−1.682[c]	−1.355[c]	−1.871[c]
	(0.175)	(0.229)	(0.191)	(0.165)	(0.208)	(0.188)	(0.209)	(0.231)	(0.258)
Gender		Female	Male		Female	Male		Female	Male
N	3588	1674	1914	3588	1674	1914	3588	1674	1914
pseudo R^2	0.113	0.111	0.091	0.070	0.076	0.057	0.088	0.100	0.078

[a] Denotes $p < 0.10$.

[b] Denotes $p < 0.05$.

[c] Denotes $p < 0.01$.

mostly impairs correct reasoning, leading to a greater likelihood of wrong (random) answers rather than intuitive ones. Furthermore, given that participants have only 30 s to read and respond to each question, there is a possibility they might not read carefully, which could lead to confusion. This could result in a higher level of non-intuitive incorrect answers under the TP treatment. This possible explanation is supported by Goeschl and Lohse (2018), where the authors find that TP leads participants to be more prone to confusion (or randomness). Thus, TP may lead to providing more non-intuitive incorrect answers instead of intuitive answers due to confusion. Another possibility is that the intuitive answers of the CRT do not really capture intuition. Indeed, there is evidence that while correct answers in the CRT are a reliable measure of deliberation, intuitive answers may not be a reliable measure of intuition (Pennycook et al., 2016). In particular, it has been shown that correct answers in the CRT correlate with the Need for Cognition (Cacioppo & Petty, 1982), a scale that measures the tendency of individuals to engage in complex cognitive tasks, while intuitive answers do not correlate with the Faith on Intuition (Epstein et al., 1996), a scale that measures the individuals' tendency in engaging in effortless and intuitive tasks. To distinguish between these two possible interpretations, it seems worth exploring the construction of an alternative behavioral measure of actual intuitive decisions.

Furthermore, one might wonder whether our results are driven by the fact that numeracy correlates with CRT scores. However, we can reasonably dismiss concerns about such potential confounding effects because our sample is well-balanced across treatments in terms of gender, age, student status, employee status, and level of education.

Additionally, considering the power of our analysis, we cannot exclude the possibility that the cognitive manipulations that appear to have no statistically significant effect in some cases may actually have a small effect size. Therefore, it would be valuable to conduct a similar study with increased sensitivity to detect smaller effect sizes, such as around 10%. Furthermore, if our novel approach proves to be successful, it could be employed to test the effectiveness of other cognitive manipulations that have been implemented to induce reliance on deliberation and intuition. Examples of such manipulations include cognitive load (Gilbert & Hixon, 1991; Gilbert et al., 1993; Schulz et al., 2014; Swann et al., 1990), conceptual priming (Cappelen et al., 2013; Capraro, Everett, & Earp, 2019; Rand et al., 2012; Shenhav et al., 2012), motivated delay (Bilancini, Boncinelli, Guarnieri & and Spadoni, 2023; Takemura, 1993; Bilancini, Boncinelli, & Spadoni, 2023; Bilancini et al., 2022), and ego depletion (Achtziger et al., 2018; Baumeister, 2002; Baumeister et al., 1998; Muraven & Slessareva, 2003; Muraven et al., 1998; Wang et al., 2017).

CRediT authorship contribution statement

Ennio Bilancini: Writing – original draft, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization. **Leonardo Boncinelli:** Writing – original draft, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization. **Tatiana Celadin:** Writing – review & editing, Writing – original draft, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare no competing interests.

Data availability

Data will be made available on request.

Acknowledgments

The authors gratefully acknowledge financial support from the Italian Ministry of Education, University and Research (MIUR) through the PRIN project Co.S.Mo.Pro.Be. "Cognition, Social Motives and Prosocial Behavior" (grant n. 20178293XT) and from the IMT School for Advanced Studies Lucca, Italy through the PAI project Pro.Co.P.E. "Prosociality, Cognition, and Peer Effects". ChatGPT was used to polish the text.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.socec.2024.102273.

References

Achtziger, A., Alós-Ferrer, C., & Wagner, A. K. (2018). Social preferences and self-control. *Journal of Behavioral and Experimental Economics*, *74*, 161–166.

Albano, G. L., Cipollone, A., Di Paolo, R., Ponti, G., & Sparro, M. (2018). Scoring rules in experimental procurement. *Journal of Behavioral and Experimental Economics*, *108*, Article 102131.

Alonso, J., Di Paolo, R., Ponti, G., & Sartarelli, M. (2018). Facts and misconceptions about 2D: 4D, social and risk preferences. *Frontiers in Behavioral Neuroscience*, *12*, 22.

Alós-Ferrer, C., & Garagnani, M. (2020). The cognitive foundations of cooperation. *Journal of Economic Behavior and Organization*, *175*, 71–85.

Andersson, O., Holm, H. J., Tyran, J.-R., & Wengström, E. (2016). Risk aversion relates to cognitive ability: Preferences or noise? *Journal of the European Economic Association*, *14*(5), 1129–1154.

Baumeister, R. F. (2002). Ego depletion and self-control failure: An energy model of the self's executive function. *Self and Identity*, *1*(2), 129–136.

Baumeister, R. F., Bratslavsky, E., Muraven, M., & Tice, D. M. (1998). Ego depletion: Is the active self a limited resource? *Journal of Personality and Social Psychology*, *74*(5), 1252.

Belloc, M., Bilancini, E., Boncinelli, L., & D'Alessandro, S. (2019). Intuition and deliberation in the stag hunt game. *Scientific Reports*, *9*(1), 14833.

Besedeš, T., Deck, C., Sarangi, S., & Shor, M. (2012). Decision-making strategies and performance among seniors. *Journal of Economic Behavior and Organization*, *81*(2), 524–533.

Bilancini, E., Boncinelli, L., & Celadin, T. (2022). Social value orientation and conditional cooperation in the online one-shot public goods game. *Journal of Economic Behavior and Organization*, *200*, 243–272.

Bilancini, E., Boncinelli, L., Guarnieri, P., & Spadoni, L. (2023). Delaying and motivating decisions in the (bully) dictator game. *Journal of Behavioral and Experimental Economics*, *107*, Article 102106.

Bilancini, E., Boncinelli, L., & Spadoni, L. (2023). Motivating risky choices increases risk taking. *Journal of Neuroscience, Psychology, and Economics*, *16*(4), 182–193.

Borghans, L., Meijers, H., & Ter Weel, B. (2008). The role of noncognitive skills in explaining cognitive test scores. *Economic Inquiry*, *46*(1), 2–12.

Bosch-Domènech, A., Brañas-Garza, P., & Espín, A. M. (2014). Can exposure to prenatal sex hormones (2D: 4D) predict cognitive reflection? *Psychoneuroendocrinology*, *43*, 1–10.

Branas-Garza, P., Kujal, P., & Lenkei, B. (2019). Cognitive reflection test: Whom, how, when. *Journal of Behavioral and Experimental Economics*, *82*, Article 101455.

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology*, *42*(1), 116.

Campitelli, G., & Labollita, M. (2010). Correlations of cognitive reflection with judgments and choices. *Judgment and Decision Making*, *5*(3), 182–191.

Cappelen, A. W., Sø rensen, E. Ø., & Tungodden, B. (2013). When do we lie? *Journal of Economic Behavior and Organization*, *93*, 258–265.

Capraro, V. (2017). Does the truth come naturally? Time pressure increases honesty in one-shot deception games. *Economics Letters*, *158*, 54–57.

Capraro, V. (2024). The dual-process approach to human sociality: Meta-analytic evidence for a theory of internalized heuristics for self-preservation. *Journal of Personality and Social Psychology*.

Capraro, V., & Cococcioni, G. (2015). Social setting, intuition and experience in laboratory experiments interact to shape cooperative decision-making. *Proceedings of the Royal Society B: Biological Sciences*, *282*(1811), Article 20150237.

Capraro, V., & Cococcioni, G. (2016). Rethinking spontaneous giving: Extreme time pressure and ego-depletion favor self-regarding reactions. *Scientific Reports*, *6*(1), 1–10.

Capraro, V., Everett, J. A., & Earp, B. D. (2019). Priming intuition disfavors instrumental harm but not impartial beneficence. *Journal of Experimental Social Psychology*, *83*, 142–149.

Capraro, V., Schulz, J., & Rand, D. G. (2019). Time pressure and honesty in a deception game. *Journal of Behavioral and Experimental Economics*, *79*, 93–99.

Cueva, C., Iturbe-Ormaetxe, I., Mata-Pérez, E., Ponti, G., Sartarelli, M., Yu, H., & Zhukova, V. (2016). Cognitive (IR) reflection: New experimental evidence. *Journal of Behavioral and Experimental Economics*, *64*, 81–93.

Cummins, D. D., & Cummins, R. C. (2012). Emotion and deliberative reasoning in moral judgment. *Frontiers in Psychology*, *3*, 328.

Epstein, S., Pacini, R., Denes-Raj, V., & Heier, H. (1996). Individual differences in intuitive–experiential and analytical–rational thinking styles. *Journal of Personality and Social Psychology*, *71*(2), 390.

Evans, J. S. B. (1989). *Bias in human reasoning: Causes and consequences*. Lawrence Erlbaum Associates, Inc.

Evans, J. S. B., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, *8*(3), 223–241.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*(4), 25–42.

Gilbert, D. T., & Hixon, J. G. (1991). The trouble of thinking: Activation and application of stereotypic beliefs. *Journal of Personality and Social Psychology*, *60*(4), 509.

Gilbert, D. T., Tafarodi, R. W., & Malone, P. S. (1993). You can't not believe everything you read. *Journal of Personality and Social Psychology*, *65*(2), 221.

Goeschl, T., & Lohse, J. (2018). Cooperation in public good games. Calculated or confused? *European Economic Review*, *107*, 185–203.

Gunia, B. C., Wang, L., Huang, L., Wang, J., & Murnighan, J. K. (2012). Contemplation and conversation: Subtle influences on moral decision making. *Academy of Management Journal*, *55*(1), 13–33.

Holt, C. A., Porzio, M., & Song, M. Y. (2017). Price bubbles, gender, and expectations in experimental asset markets. *European Economic Review*, *100*, 72–94.

Hoppe, E. I., & Kusterer, D. J. (2011). Behavioral biases and cognitive reflection. *Economics Letters*, *110*(2), 97–100.

Lohse, J. (2016). Smart or selfish–when smart guys finish nice. *Journal of Behavioral and Experimental Economics*, *64*, 28–40.

Lohse, T., Simon, S. A., & Konrad, K. A. (2018). Deception under time pressure: Conscious decision or a problem of awareness? *Journal of Economic Behavior and Organization*, *146*, 31–42.

Muraven, M., & Slessareva, E. (2003). Mechanisms of self-control failure: Motivation and limited resources. *Personality and Social Psychology Bulletin*, *29*(7), 894–906.

Muraven, M., Tice, D. M., & Baumeister, R. F. (1998). Self-control as a limited resource: regulatory depletion patterns. *Journal of Personality and Social Psychology*, *74*(3), 774.

Palan, S., & Schitter, C. (2018). Prolific. AC—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27.

Pennycook, G., Cheyne, J. A., Koehler, D. J., & Fugelsang, J. A. (2016). Is the cognitive reflection test a measure of both reflection and intuition? *Behavior Research Methods*, *48*(1), 341–348.

Ponti, G., & Rodriguez-Lara, I. (2015). Social preferences and cognitive reflection: Evidence from a dictator game experiment. *Frontiers in Behavioral Neuroscience*, *9*, 146.

Primi, C., Morsanyi, K., Chiesi, F., Donati, M. A., & Hamilton, J. (2016). The development and testing of a new version of the cognitive reflection test applying item response theory (IRT). *Journal of Behavioral Decision Making*, *29*(5), 453–469.

Rand, D. G., Greene, J. D., & Nowak, M. A. (2012). Spontaneous giving and calculated greed. *Nature*, *489*(7416), 427–430.

Rand, D. G., Peysakhovich, A., Kraft-Todd, G. T., Newman, G. E., Wurzbacher, O., Nowak, M. A., & Greene, J. D. (2014). Social heuristics shape intuitive cooperation. *Nature Communications*, *5*(1), 1–12.

Schulz, J. F., Fischbacher, U., Thöni, C., & Utikal, V. (2014). Affect and fairness: Dictator games under cognitive load. *Journal of Economic Psychology*, *41*, 77–87.

Shenhav, A., Rand, D. G., & Greene, J. D. (2012). Divine intuition: Cognitive style influences belief in god. *Journal of Experimental Psychology: General*, *141*(3), 423.

Suter, R. S., & Hertwig, R. (2011). Time and moral judgment. *Cognition*, *119*(3), 454–458.

Swann, W. B., Hixon, J. G., Stein-Seroussi, A., & Gilbert, D. T. (1990). The fleeting gleam of praise: Cognitive processes underlying behavioral reactions to self-relevant feedback. *Journal of Personality and Social Psychology*, *59*(1), 17.

Takemura, K. (1993). The effect of decision frame and decision justification on risky choice. *Japanese Psychological Research*, *35*(1), 36–40.

Trémolière, B., & Bonnefon, J.-F. (2014). Efficient kill–save ratios ease up the cognitive demands on counterintuitive moral utilitarianism. *Personality and Social Psychology Bulletin*, *40*(7), 923–930.

Wang, Y., Wang, G., Chen, Q., & Li, L. (2017). Depletion, moral identity, and unethical behavior: Why people behave unethically after self-control exertion. *Consciousness and Cognition*, *56*, 188–198.

Wason, P. C., & Evans, J. S. B. (1974). Dual processes in reasoning? *Cognition*, *3*(2), 141–154.